PATENT ABSTRACTS OF JAPAN

(11)Publication number : 01-093795
(43)Date of publication of application : 12.04.1989

(51)Int.Cl. G10L 3/02

(21)Application number : 62-250707
(22)Date of filing : 06.10.1987

(71)Applicant : NIPPON HOSO KYOKAI <NHK>
(72)Inventor : TSUGI TORU
KUWABARA HISAO

(54) METHOD FOR CONVERTING VOCALIZATION SPEED OF VOICE

(57)Abstract:
PURPOSE: To hold the continuity of an waveform and to suppress the deterioration of sound quality by separating an input voice into vowel and consonant sections and changing vocalization speed in each section in accordance with a vocalization feature.
CONSTITUTION: The voice section and silent section of an A/D converted input voice are discriminated by an analysis part 2, the voiceless consonant section and voiced section of the input voice are discriminated and these waveforms are stored. A linear prediction coefficient and a residual waveform in the voiced section are found out, a pitch period is also found out to determine one pitch section and normalization power is defined. A vowel is separated from a voiced consonant part by using resonance frequency and the normalization power. When a control part 4 extends the length of a silent section or repeats or thins respective pitches of the voiced section by proper distribution, a vocalization speed is changed and a new pitch period string is prepared. An waveform connection part 6 connects respective parts by extending/shortening their vocalization time length based upon the new pitch period string to obtain a new voice waveform.

(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号

第2612868号

(45)発行日　平成９年(1997)５月21日　(24)登録日　平成９年(1997)２月27日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ１０Ｌ	3/02			Ｇ１０Ｌ	3/02	Ａ
	9/14				9/14	Ｂ

発明の数１(全 7 頁)

(21)出願番号　特願昭62－250707

(22)出願日　昭和62年(1987)10月６日

(65)公開番号　特開平1－93795

(43)公開日　平成１年(1989)４月12日

特許権者において、実施許諾の用意がある。

(73)特許権者　999999999
日本放送協会
東京都渋谷区神南２丁目２番１号

(72)発明者　都木　徹
東京都世田谷区砧１丁目10番11号　日本放送協会放送技術研究所内

(72)発明者　桑原　尚夫
大阪府枚方市東香里１丁目21番地25―203

(74)代理人　弁理士　谷　義一

審査官　河口　雅英

(56)参考文献　特開　昭59－82608（ＪＰ，Ａ）
特開　昭61－55698（ＪＰ，Ａ）

(54)【発明の名称】　音声の発声速度変換方法

(57)【特許請求の範囲】

【請求項１】入力音声波形から、母音区間、有声子音区間、無声子音区間、無音区間を抽出し、

前記有声子音区間と前記母音区間とで構成される有声音区間からピッチ周期を抽出することによって該有声音区間を当該ピッチの間隔で分割し、

前記母音区間および前記無音区間における発話時間長の伸縮比率を大とし、かつ前記有声子音区間および前記無声子音区間の前記伸縮比率を小とする前記各々の区間の前記伸縮比率を定め、

前記母音区間および前記有声子音区間では前記定められた伸縮比率に基づき前記ピッチ間隔で波形の間引または繰り返しをすることによって発声時間長を伸縮し、

前記無声子音区間および前記無音区間では前記定められた伸縮比率に基づき当該区間毎に発声時間長の伸縮を行

なった後前記各々の区間を接続して新たな音声波形とすることを特徴とする音声の発声速度変換方法。

【発明の詳細な説明】

[産業上の利用分野]

本発明は、放送、映画、音楽等において、人間の音声を処理する場合の発声速度を制御する音声の発声速度変換方法に関する。

[発明の概要]

本発明は人の音声を一時記録し、その発生速度を変化させて、再び音声として出力する技術に関するもので、

入力音声をA/D変換した後、有声音区間についてそのピッチ周波数を抽出して各ピッチ間隔で分割し、その内おもに定常母音区間についてピッチ単位で間引きまたは繰り返しを行うと共に、無音区間、無声子音区間についても間引きまたは繰返しを行って接続し、これをD/A変

換することにより、

原音声の音韻性や自然性を良好に保ったまま、発声速度を自由に変換できるようにする方法である。

［従来の技術］

この種の技術としては、古典的な例として音声をアナログテープレコーダに録音し、再生スピードを変化させる方法がある。この場合、発声速度のみならず、ピッチ周波数やホルマント周波数も一様に変化する。すなわち、再生スピードを録音時のＲ倍にすると、発声速度がＲ倍になると共に、ピッチおよびホルマント周波数も全てＲ倍となる。ここで、ピッチ周波数はその全体的な変化によって音声の高低を決定し、局所的な変化によって、アクセント等、音声の抑揚を決定するものである。また、ホルマント周波数は音声の個人性や音韻性を定めるものである。

これに対し、Ｒ倍になったピッチおよびホルマント周波数を元に戻すには、BBDなどを用いてクロック周波数Ｆで取込んだ音声波形を、F/Rなるクロック周波数で読出せばピッチおよびホルマント周波数が1/R倍となりもとに戻る。ただし、BBDに取込む前に、適当な時間窓と周期を用いて波形を間引いたり、繰り返したりして、過不足のないようにする。

また、デジタル信号処理である、分析・合成法を用いる方式も提案されている。分析によって得られた調音パラメータと残差波形を、時間的に適当な単位で間引いたり、繰返しながら合成すれば、ピッチおよびホルマント周波数には変化を与えずに発声速度を制御することができる。

［発明が解決しようとする問題点］

しかしながら、テープレコーダの再生スピードを変化させるだけの方法は簡単ではあるが、ピッチやホルマント周波数も変化してしまう。ピッチやホルマント周波数が変化すると、個人性に影響があり、更に変化量が多い場合には音韻性が劣化し、非人間的な声となる。

またピッチやホルマント周波数を元に戻す方式においても、その処理単位が、ブロック単位であるため、波形の連続性を完全に保つことが難しく、音質劣化が著しい。

さらに、分析・合成方法においても、出力音声がパラメータ制御による合成音であるためある程度の音質劣化は避けられない。

また、従来の方式では、処理が全ての区間で一様であるが、実際の音声では子音の種類によってはその持続時間が発声速度に殆ど依存せず、この部分を母音区間と同じ比率で時間伸縮したのでは、会話音声としての自然性が劣化する。

さらにｔやｋのような破裂性の子音は持続時間が短いので、ブロック単位で間引いた場合に消失する場合がある。

そこで、本発明の目的は上述した従来の問題点を解消

し、間引きや繰り返しの単位をピッチ単位とすることで波形の連続性を保ち、かつ原音声の波形をそのまま用いることで音質の劣化を防ぐことを可能とする音声の発声速度変換方法を提供することにある。

本発明の他の目的は母音区間、有声子音区間、無声子音区間、無音区間を別々の比率で時間伸縮し、音声としての自然性を維持することが可能な音声の発声速度変換方法を提供することにある。

［問題点を解決するための手段］

そのために本発明では入力音声波形から、母音区間、有声子音区間、無声子音区間、無音区間を抽出し、有声子音区間と母音区間とで構成される有声音区間からピッチ周期を抽出することによって有声音区間をピッチの間隔で分割し、母音区間および無音区間における発話時間長の伸縮比率を大とし、かつ有声子音区間および無声子音区間の伸縮比率を小とする各々の区間の伸縮比率を定め、母音区間および有声子音区間では定められた伸縮比率に基づきピッチ間隔で波形の間引または繰り返しをすることによって発声時間長を伸縮し、無声子音区間および無音区間では定められた伸縮比率に基づき区間毎に発声時間長の伸縮を行なった後各々の区間を接続して新たな音声波形とすることを特徴とする。

［作　用］

以上の構成によれば、入力音声を母音区間、有声子音区間、無声子音区間、無音区間に分離し、それぞれの区間毎に人間の発声特徴に応じた変換方法を用いて発声速度を変換する。

すなわち、有声音区間では音声の間引きや繰り返しの単位をピッチ単位とし、かつ原音声の波形をそのまま用いる。

また、子音区間においても、それぞれの子音の性質により伸縮の方式を切替える。

［実施例］

以下、図面に示す実施例に基づき本発明を詳細に説明する。

第１図は、本発明の一実施例に係る発声速度変換システムのブロック図を示す。図において、２は分析部、４は制御部、６は波形接続部をそれぞれ示し、各部は電子計算機内に構成され、ROM,RAMあるいはメモリディスク等のメモリを併用しながら発声速度変換の処理が実行される。

A/D変換されて標本化された音声波形は分析部２へ入力し、有音と無音および有声音と無声音の判別、さらには有声音については線形予測分析がなされ、ピッチ周期、予測係数、共振周波数、共振の帯域幅が求められる。

次に、制御部４においては、発声速度を変更し、波形接続部６では発声時間長を伸縮して波形の接続を行なう。

上述した一連の発声速度変換の処理を終了すると、合

成された音声波形をD/A変換して出力音声とする。

上記各部における処理の詳細を第２図に示すフローチャートを参照しながら説明する。

変換ビット数12bit,標本化周波数15kHzでA/D変換された音声は、まず、分析部２において、ステップS1で音声パワーの有無に基づいて有音区間と無音区間の判別が行われる。次にステップS2では有音区間の標本値に対してPARCOR分析と零交さ分析とを行い、無声子音区間と有声音区間との判別を行う。これは、１次のPARCOR係数を参照して入力周波数の高域成分の割合を調べたり、零交さ数を調べることによって行なう。すなわち、無声音のエネルギーは高周波領域にまで分布しているので、高域成分の割合および高周波になると多くなる零交さ数を調べることによって無声子音と有声音とを判別する。なお、PARCOR分析と零交さ分析の両方を用いて判別を行なうのは、判別を確実なものとするためである。

上記ステップS1およびS2で判別された無音区間の時間および無声子音区間の波形は、それぞれステップS15およびS16においてそのままRAMあるいはメモリディスク等に記憶される。

次に、ステップS3では有声音区間における音声波形の標本値を音声の生成モデルに基づくいわゆる声道逆フィルタに通すことによって線形予測分析を行なう。この線形予測分析によって線形予測係数と残差波形を得る。得られた残差波形はステップS17においてRAMあるいはメモリディスク等に記憶される。

ステップS4ではステップS3で得られた残差波形の相間における周期と原音声波形のピークの間隔とから仮のピッチ周期を求める。

次に、ステップS5においては、第３図に示すように波形のレベルが急に大きくなる点の直前をピッチの開始点とし、上記で求めたピッチ周期に基づき次のピッチの開始点の１標本手前を終了点として１つのピッチ区間を定める。

ステップS6では上記で求めた１ピッチ区間の中間点を分析窓の中心として、20msec程度の窓掛けを行なう。この窓掛けにより有限個の標本値による短時間スペクトル分析が可能となり、この窓掛けデータを基に再び線形予測分析を行なう。すなわち、標本値の窓掛けを行なったデータを基に相関関数を求めることによって、線形予測係数$\alpha_1 \sim \alpha_p$を算出する。ここで、pは線形予測分析の次数であり、一般に男性の声に対してはp＝14、女性の声に対してはp＝10程度を用いる。

さらに、ステップS18で、以下に示す（１）式を満足するzの根$Z_1 \sim Z_p$を求め、各々の根Z_iに対応して（２），（３）式により共振周波数F_iとその帯域幅B_iを求める。

$$1+\alpha_1 z^{-1}+\alpha_2 z^{-2}+\cdots+\alpha_p z^{-p}=0 \qquad (1)$$

$$F_i=Fs/(2\pi)\cdot\arg(z_i) \quad [Hz] \qquad (2)$$

$$B_i=Fs/\pi\cdot|\log(|z_i|)| \quad [Hz] \qquad (3)$$

なおFsは音声の標本化周波数である。

また、ステップS7はこの１ピッチ区間内のサンプル値の自乗和をピッチ区間長で割った値を正規化パワーと定義し、ピッチ区間の長さと共にRAMあるいはメモリディスク等に記録する。

処理区間を１ピッチ分だけ後へずらし、上述した一連の処理を行い、これらの操作を有声区間が終るまで繰返す。

（２）式で求めた共振周波数の時間軌跡は、定常母音部では連続的にかつ緩やかに変化するが、有声子音部では不安定に変化しかつ帯域幅は母音部よりも広い。また正規化パワーの時間軌跡においては有声子音部で一時的かつ急激な減少が起こることが多い。そこで、ステップS8では、これらの特徴を用いて、母音部と有声子音部を分離し、各ピッチ毎にその情報をRAMあるいはメモリディスク等に記録する。

制御部４では、分析部２において得られた、無音区間長や一連のピッチ周期を基に、適当な配分により無音区間長を伸縮したり、有声区間の各々のピッチを繰返すかまたは間引くことにより、発話の時間長即ち発声速度が変更された新しいピッチ周期列を作る。

ここで分析部２において次のような結果が得られたとする。

全発声時間長	T_{all}
母音部分の時間長の総和	T_v
有声子音部分の時間長の総和	T_{cv}
無声子音部分の時間長の総和	T_{cn}
無音部分の時間長の総和	T_s

ただし

$$T_{all}=T_v+T_{cv}+T_{cn}+T_s \qquad (4)$$

ここで発声速度をR倍にしたければ、T_{all}を1/R倍にすればよい。

ところが、実際の音声では、発声速度が変化してもT_{cn}やT_{cv}はあまり変化せず、主にT_sやT_vが変化する。そこで、T_sとT_vについては１の重みで、T_{cn}とT_{cv}についてはw（ただしw＜１）の重みでその長さを変更し、その和T'_{all}がT_{all}の1/R倍になるようにする。すなわちステップS9において、変更後の各部の時間長を次のようにする。

$$T'_{all}=\gamma_0\cdot T_{all} \qquad (5)$$

$$T'_{v}=\gamma_1\cdot T_v \qquad (6)$$

$$T'_{cv}=\gamma_2\cdot T_{cv} \qquad (7)$$

$$T'_{cn}=\gamma_2\cdot T_{cn} \qquad (8)$$

$$T'_{s}=\gamma_1\cdot T_s \qquad (9)$$

ただし

$$\gamma_0=1/R \qquad (10)$$

7　　8

$$\gamma_1=\frac{\gamma_0\cdot T_{all}-(1-w)\cdot(T_{cv}+T_{cn})}{T_v+T_s+w\cdot(T_{cv}+T_{cn})} \quad (11)$$

$$\gamma_2=\frac{\gamma_0\cdot w\cdot T_{all}+(1-w)\cdot(T_v+T_s)}{T_v+T_s+w\cdot(T_{cv}+T_{cn})} \quad (12)$$

ここでwの値は、0.3～0.5程度とする。

波形接続部６では制御部４で決定された比率により各部分の発声時間長を伸縮して接続する。

母音区間、有声子音区間においてそれぞれの発声時間長をγ_1倍、γ_2倍にするには、以下のように適当な割合でピッチ単位の波形を適宜間引くかまたは繰り返して接続する。

すなわち、ステップS10およびS11で、ある母音区間または有声子音区間の発声時間長をγ倍するとして、γ＞１ならば、1/（γ－１）ピッチにつき１ピッチの割合で同じピッチ波形を繰返し、γ＜１ならば、1/（１－γ）ピッチにつき１ピッチの割合で間引く。第４図にγ＝1.5、およびγ＝0.667の場合の例を示す。同図から明らかなように、γ＝1.5の場合は２ピッチに１回ピッチ区間２および４を繰り返えす。また、γ＝0.667の場合、３ピッチに１回ピッチ区間３および６を間引く。

なお、有声子音区間のうち原音声の区間長が25msec以下のものについては流音／γ／の可能性が高く、この区間の長さは発声速度には殆ど依存しないので伸縮は行わない。

このようにすれば、概ね原音声のγ倍の発声時間長とすることができ、かつ聴感的にも違和感がない。

なお、一般的にピッチ区間を間引くかまたは繰返した波形においては、あるピッチ区間の終了点と次のピッチ区間の開始点の間は不連続であるので、接続点の前後数サンプルのデータを用いて最小自乗法により３次曲線を用いた近似を行い、連続的に接続する。

無声子音区間においてはステップS12で原音声の区間長Ｌが60msecより短いものについては破裂性または破擦性の子音の可能性が高いので、それ自身の伸縮は行わない。

Ｌが60ミリ秒より大きいものについてはγ_2＜１ならば区間の開始点および終了点から中間点に向かって、それぞれＬ・（１－γ_2）/2に相当する長さを省く。２≧γ_2＞１ならば中間点の前後Ｌ・（γ_2－１）に相当する長さの波形を切り出し原波形の中間点の間に挿入する。この様子を第５図に示す。γ_2＞２の場合は、全区間を繰返す操作を適宜加える。

無音区間においては、ステップS13で、基本的には無条件にその区間長をγ_1倍して新たな区間長とするが、無声子音の直後の30ミリ秒以下の無音部は、無声破裂子音の気音部の可能性が高いので例外としてその長さを不変とすると共に、無声子音の直前の無音部を短くする場合には30ミリ秒以下にならないように制限する。

なお、以上の処理で各部分に生じた伸縮時間長の誤差は、それぞれの区間の近傍の無音区間または母音区間の長さを伸縮して修正する。

ひとつの区間の処理が終了したならば、ステップS14において、その開始部および終了部に１ミリ秒程度の立上がりおよび立下がりの窓をかけ、前の区間と接続し、次の区間の処理に移る。

なお、長時間にわたる連続音声の全発声時間長を基に処理を行うのは困難であるので、100～200ミリ秒前後の比較的長い無音区間を検出したならば、その中間点までをひとつのブロックと考え、まずこの１ブロックの中で上記の一連の時間伸縮処理を行った後、つぎのブロックの処理に移る。ただし、原音声が比較的早口の場合には、ブロック分割を判断するための無音区間長を50ミリ秒程度に狭めた方がよい。

最終的に合成された音声をD/A変換して、出力音声とする。

なお、分析部２における、ピッチ周波数抽出法や、有声／無声判別法、有声子音抽出法などは、ここで述べたものに限らず、それらが精度良く抽出できる方法なら何でも良い。

［発明の効果］

以上説明したように、本発明によれば予め入力音声を母音区間、有声子音区間、無声子音区間、無音区間に分離し、それぞれの区間毎に人間の発声の特徴に応じた変換方法を用いて発声速度を換えるので、音声としての自然性が高い。

また、有声音区間では音声の間引きや繰返しの単位をピッチ単位とすることで波形の連続性を保ち、かつ原音声の波形をそのまま用いることで音質の劣化が殆どない。

さらに子音区間においても、それぞれの子音の性質により伸縮の方式を切替えることができるので、持続時間の短いものが脱落することなどもなく、明瞭度の低下を最小限に抑えることができる。

【図面の簡単な説明】

第１図は本発明の一実施例に係るシステムのブロック図、

第２図は本発明の一実施例を示すフローチャート、

第３図は実施例におけるピッチ区間の定め方を説明するための波形図、

第４図は実施例における波形の繰り返しおよび間引きを説明するための波形図、

第５図は実施例における無声子音部の波形の伸縮を説明

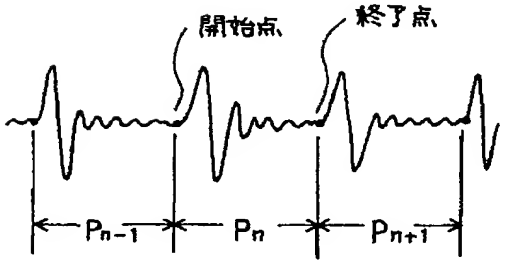
9　　10

するための波形図である。
２……分析部、
＊４……制御部、
＊　６……波形制御部。

【第１図】

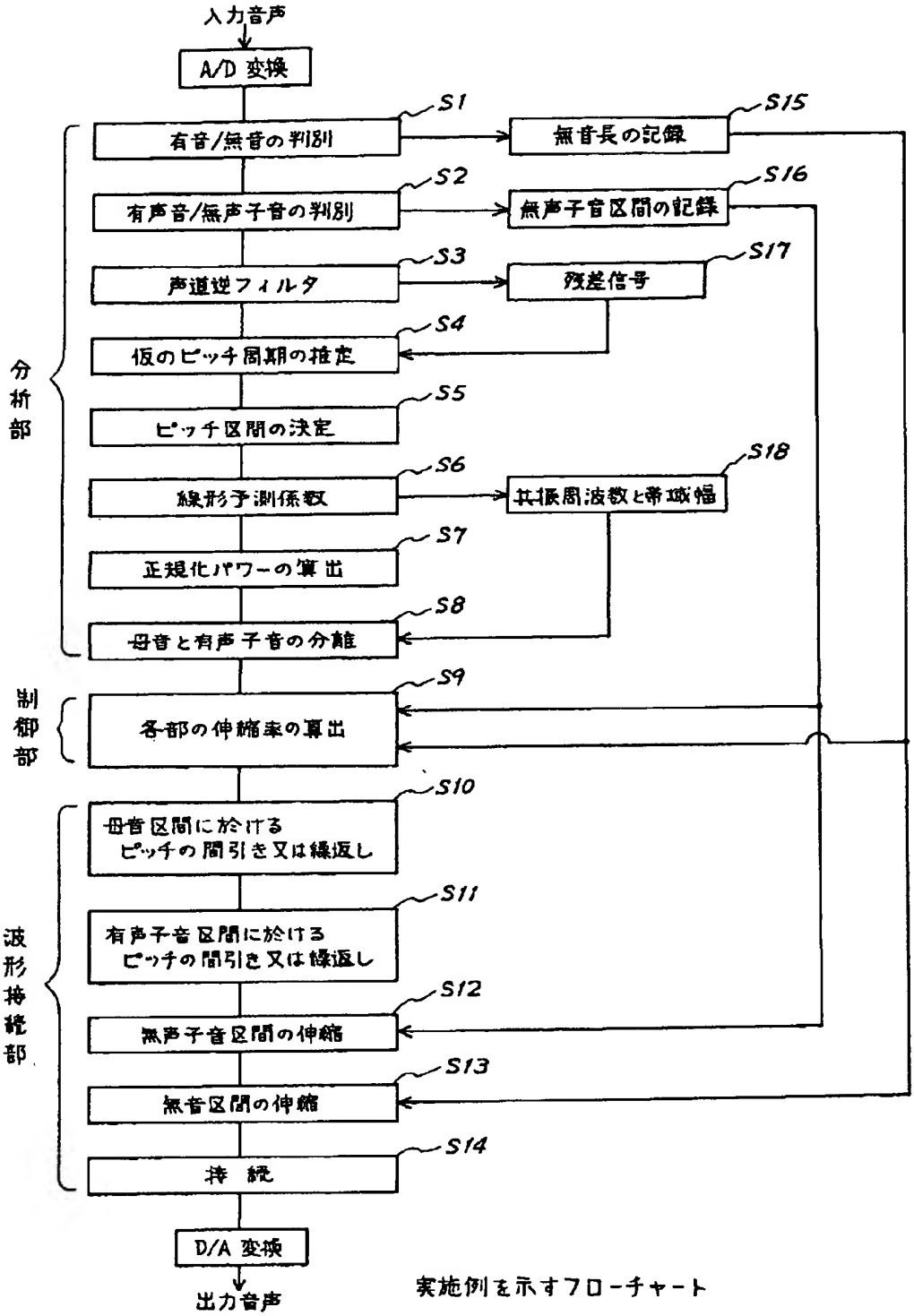
実施例に係るシステムのブロック図

【第３図】



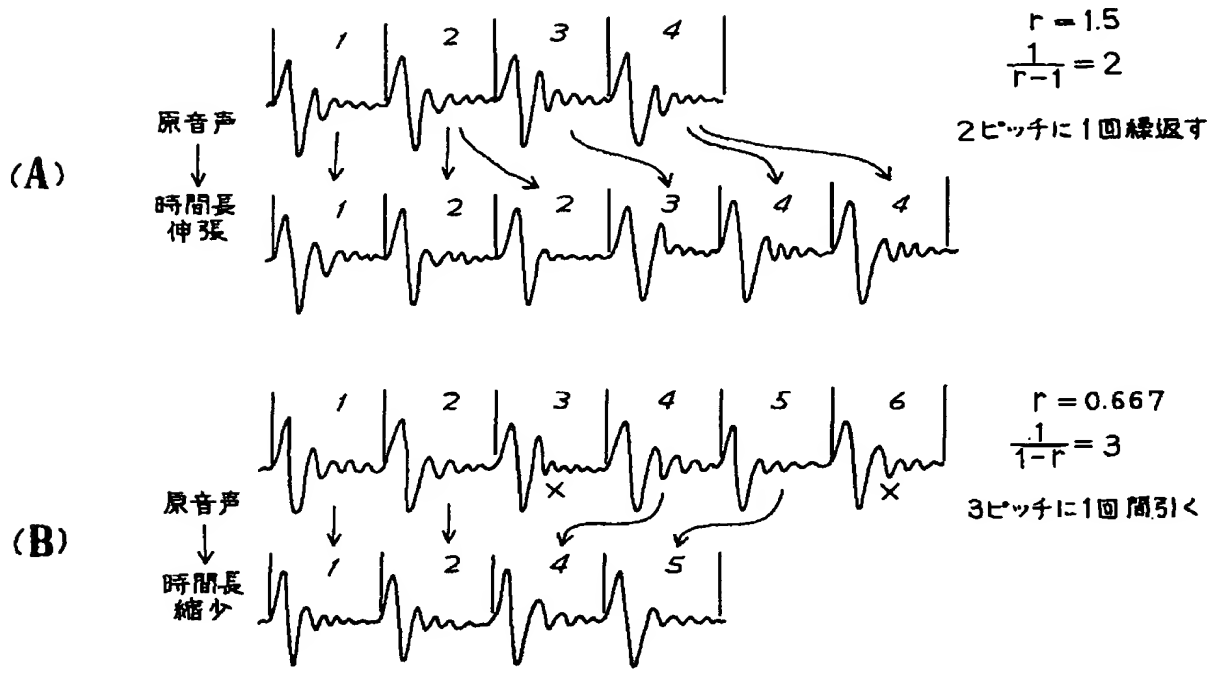
実施例の波形図

【第２図】



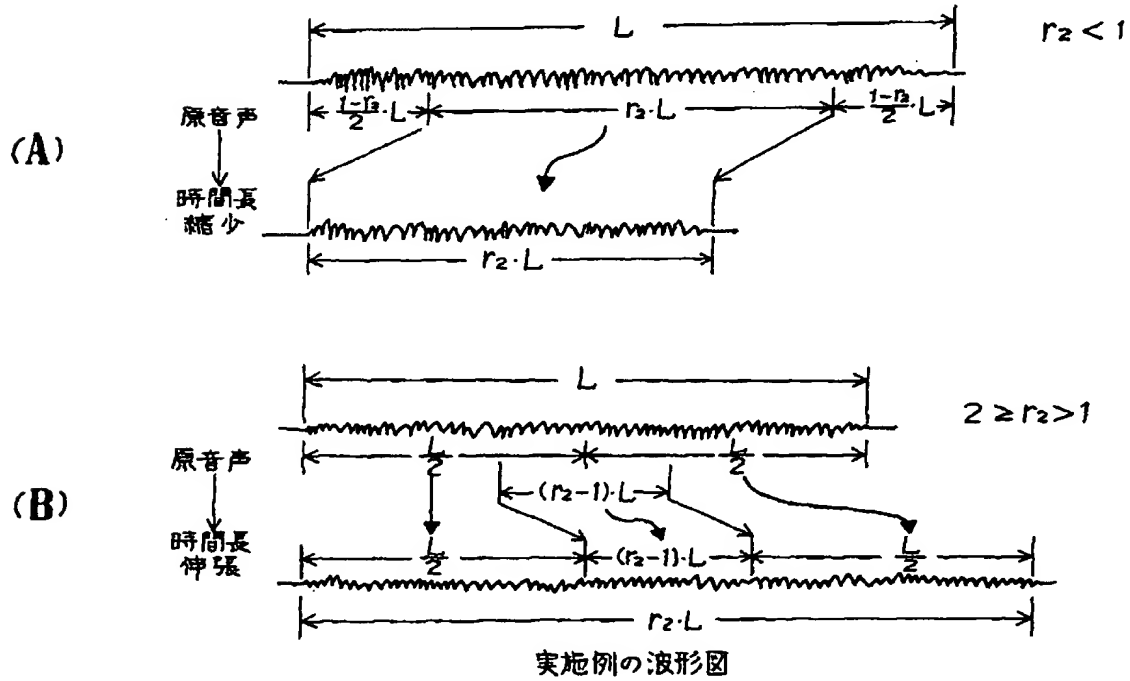
実施例を示すフローチャート

【第４図】



実施例の波形図

【第５図】



実施例の波形図